Homework 2

Anna Pauxberger

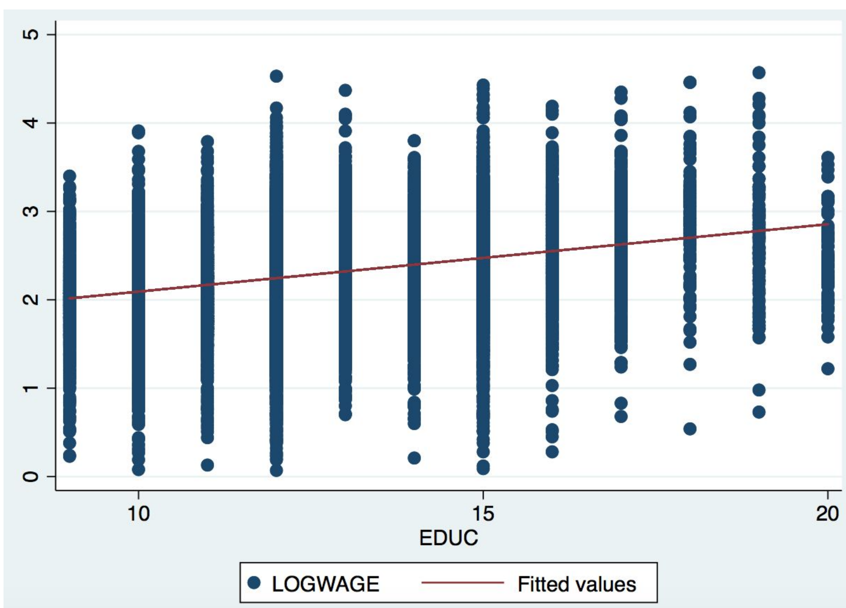
03 November 2018

Q1) Using the link (<u>http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataSets.htm</u>), download the data used in Koop and Tobias's (2004) study of the relationship between wages and education, ability, and family characteristics. Their data set is a panel of 2,178 individuals with a total of 17,919 observations. Extract the first observations for the first 15 individuals in the sample.

Let X1 equal a constant, education, experience, and ability (the individual's own characteristics). Let X2 contain the mother's education, the father's education, and the number of siblings (the household characteristics). Let y be the log wage.

a. Compute the least squares regression coefficients in the regression of **y **on X1. Report and interpret the coefficients.

The least squares regression minimizes the squares of the distance between the line of best fit and the actual data points. The visualization below shows the line of best fit for log wage when regressed on education.



```
. reg $y_list $x1_list

      Source |       SS           df       MS      Number of obs   =       109
-------------+----------------------------------   F(3, 105)       =     18.09
       Model |  7.79915316         3  2.59971772   Prob > F        =    0.0000
    Residual |  15.0874719       105  .143690209   R-squared       =    0.3408
-------------+----------------------------------   Adj R-squared   =    0.3219
       Total |  22.8866251       108  .211913195   Root MSE        =    .37906

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1283769   .0250904     5.12   0.000     .0786272    .1781265
    potexper |   .0442788   .0087591     5.06   0.000     .0269111    .0616464
     ability |  -.1481151   .0535166    -2.77   0.007    -.2542285   -.0420016
       _cons |   .3292917    .342301     0.96   0.338     -.349428    1.008011
------------------------------------------------------------------------------
```

Both this and the following regression were run with the robust option to account for potential heteroskedasticity in the error term, meaning that errors are not distributed in the same way across all x-variables. The robust option ensures correct, robust standard errors. As visible in the appendix, the robust standard errors differ slightly from the normal standard errors. To show the model and residual square sums, the regression output without the robust option was chosen.

Coefficients interpretation: We are regressing education, experience and ability on log wage. In this log-linear model, we can interpret a one unit change of any x variable leading to a 100*(coefficient)% change in wage. For the above regression, all coefficients (except for the constant) are statistically significant at an alpha level of 0.01 (p-value below 0.01). The 95% confidence intervals are neither particularly small nor particularly large for all coefficients, such as education's coefficient of 0.12 with a confidence interval of [0.07, 0.17].

educ	0.128	expectedly, a 1 year increase in education is associated with a 12.8% increase of wage
potexper	0.044	expectedly, a 1 year increase in potential experience is associated with a 4% increase of wage (potential experience is defined as Age -Education-5 in the paper footnote on page 834)
ability	-0.148	unexpectedly, a 1 unit increase of ability is associated with a 1% decrease in wage (ability is measured via a 10-component cognitive test)
constant	0.329	y-intercept which shifts the regression vertically upwards by 0.33 (though *not* statistically significant)

b. Compute the least squares regression coefficients in the regression of **y **on X1 and X2. Report and interpret the coefficients.

```
. reg $y_list $x1_list $x2_list

      Source |       SS           df       MS      Number of obs   =       109
-------------+----------------------------------   F(6, 102)       =      9.81
       Model |  8.37678951         6  1.39613159   Prob > F        =    0.0000
    Residual |  14.5098356       102  .14225329    R-squared       =    0.3660
-------------+----------------------------------   Adj R-squared   =    0.3287
       Total |  22.8866251       108  .211913195   Root MSE        =    .37716

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1359242   .0254552     5.34   0.000     .0854339    .1864146
    potexper |   .0475567   .0089154     5.33   0.000      .029873    .0652404
     ability |   -.074606   .0857221    -0.87   0.386    -.2446353    .0954233
    mothered |   .0111151   .0401131     0.28   0.782    -.0684491    .0906792
    fathered |  -.0342395    .023343    -1.47   0.146    -.0805402    .0120612
    siblings |   .0396161   .0550796     0.72   0.474     -.069634    .1488662
       _cons |   .3815602   .5607023     0.68   0.498    -.7305901    1.493711
------------------------------------------------------------------------------
```

Coefficient interpretation: Combining the x variable lists leaves only education and potential experience with statistical significance at alpha=0.01. All other variables indicate a p-value of above 0.01. The 95% confidence intervals remain similar to the previous regression, such as with education's coefficient of 0.136 having a 95% confidence interval of [0.085, 0.186].

educ	0.136	(increased from 0.128 to 0.136) a 1 year increase in education leads to a 13.5% increase of wage
potexper	0.047	(increased from 0.044 to 0.047) expectedly, a 1 year increase in potential experience leads to a 5% increase of wage

c. Compute the R-squared for the the regression of **y **on X1 and X2 manually using the SSE and SST from the output. Repeat the computation for the case in which the constant term is omitted from X1. What happens to R-squared?

```
reg $y_list $x1_list $x2_list
```

Source	SS		df	MS
Model	8.37678951	SSR	6	1.39613159
Residual	14.5098356	SSE	102	.14225329
Total	22.8866251	SST	108	.211913195

Number of obs	=	109	
F(6, 102)	=	9.81	
Prob > F	=	0.0000	
R-squared	=	0.3660	R^2
Adj R-squared	=	0.3287	
Root MSE	=	.37716	

The R-squared value, also called the coefficient of determination, roughly indicates how well the line of best fit fits the data. In this case it is computed by Stata to be 0.3660. This indicates that 36.6% of the variation in y can be explained by the x variables. While the interpretation of R^2 depends on the context and has limitations, such as being manipulatable by adding more explanatory variables, I would argue this is a rather low R^2 value and our model is not good at fitting the data.

Manual computation:

SSE is the sum of squares due to error, which means it measures the variation in the error term. Adding meaningful variables should reduce this error. SSR is the sum of squares due to regression, which is the variation in y that can be explained by the regression itself. SST is the total sum of squares, and is a sum of SSE and SSR.

$$SST = SSR + SSE$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{14.509}{8.377} = 1 - 0.6339 = 0.3660$$

Manual computation via Stata

```
. disp e(mss) / (e(mss)+e(rss))
.36601244
```

```
. reg $y_list $x1_list $x2_list, noconstant
```

Source	SS	df	MS
Model	622.640388	6	103.773398
Residual	14.5757111	103	.141511758
Total	637.216099	109	5.84601926

Number of obs	=	109
F(6, 103)	=	733.32
Prob > F	=	0.0000
R-squared	=	0.9771
Adj R-squared	=	0.9758
Root MSE	=	.37618

logwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1456608	.0209986	6.94	0.000	.104015	.1873066
potexper	.0487579	.0087161	5.59	0.000	.0314716	.0660443
ability	-.0708643	.0853223	-0.83	0.408	-.2400809	.0983524
mothered	.024648	.0347456	0.71	0.480	-.0442616	.0935577
fathered	-.0303609	.0225774	-1.34	0.182	-.0751378	.0144161
siblings	.049938	.0528117	0.95	0.347	-.0548015	.1546775

```
. disp e(mss) / (e(mss)+e(rss))
.97712595
```

By exerting the stata option ‘noconstant’, the constant (“_cons”) has been removed from the regression, and the R^2 value increases to 98% claiming to explain almost all of the variation of y with x-variables.

d. Compute the adjusted R-squared for the full regression including the constant term. Interpret your results. Do we need the constant term?

Even though it is tempting to disregard the constant term to achieve a higher R^2 result, this is misleading since R^2 is based on Sum Squared Total (SST), which is calculated via the sum of (yi-y_bar)^2. Including a constant or intercept, y_bar is is the mean of all yi. Without an intercept, however, y_bar is taken as 0, thus it will skew the R^2 results to be very close to 1.

```
. reg $y_list $x1_list $x2_list
```

Source	SS	df	MS			
Model	8.37678951	6	1.39613159	Number of obs	=	109
Residual	14.5098356	102	.14225329	F(6, 102)	=	9.81
				Prob > F	=	0.0000
				R-squared	=	0.3660
				Adj R-squared	=	0.3287
Total	22.8866251	108	.211913195	Root MSE	=	.37716

logwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1359242	.0254552	5.34	0.000	.0854339	.1864146
potexper	.0475567	.0089154	5.33	0.000	.029873	.0652404
ability	-.074606	.0857221	-0.87	0.386	-.2446353	.0954233
mothered	.0111151	.0401131	0.28	0.782	-.0684491	.0906792
fathered	-.0342395	.023343	-1.47	0.146	-.0805402	.0120612
siblings	.0396161	.0550796	0.72	0.474	-.069634	.1488662
_cons	.3815602	.5607023	0.68	0.498	-.7305901	1.493711

When adding a new variable, the adjusted R^2 only increases the R^2 value when the improvement is larger than random improvement by chance. This should counteract the original R^2's problem of improving the value simply by adding more variables regardless of whether they actually improve the model, and prevent displaying a higher value for a model that is overfitted (too well trained on the training set), so that it will perform really poorly when tested with new data.

Manual calculation (Adjusted R2, 2018)

$$R^2{}_{adj} = 1 - [\frac{(1-R^2)\,(n-1)}{n-k-1}]$$

$$with\ n = number\ of\ samples,\ k = number\ of\ explanatory\ variables$$

$$R^2{}_{adj} = 1 - [\frac{(1-R^2)\,(n-1)}{n-k-1}] = 1 - [\frac{(1-0.366)\,(109-1)}{109-5-1}] = 1 - [\frac{(0.634)\,(108)}{103}] = 1 - 0.6647 = 0.33$$

$$with\ n = 109,\ k = 5$$

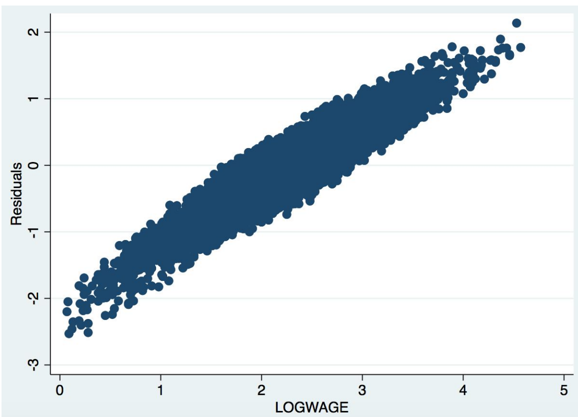
Manual calculation using Stata

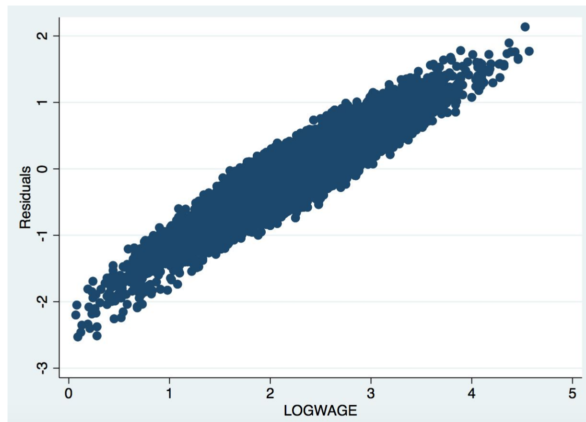
```
. gen n = e(N)

. gen r_2 = e(r2)

. gen k = 6

. disp 1-((1-r_2)*(n-1) / (n-k-1))
.32871907
```

I don't think that the adjusted R^2 term should play a role in the decision whether or not to include a constant term, since the adjusted R^2 metric does not include the constant term in its calculations. It penalizes adding explanatory variables and excludes the constant coefficient. The lack of statistical significance for any of the terms in the regression speaks for leaving out the intercept. As mentioned above, R^2 and consequently adjusted R^2 values are unreliable for computations that leave out the constant term, which speaks for it not being left out.

e. Are any of the classical assumptions violated in part a or part b? Refer to the assumptions MR1, MR2, MR5, and MR6.

MR1, the linearity and additivity assumption holds, as y is a linear function of the explanatory variables x. MR2, the assumption that errors have an expected value of zero, seems not to hold as the graphs show a slight skewness towards negative residuals. However, when taking the mean of the residuals they are very close to zero, thus we can conclude the assumption holds.






```
. summarize residuals
```

Variable	Obs	Mean	Std. Dev.	Min	Max
residuals	17,919	1.07e-10	.4802517	-2.528907	2.136589

```
. summarize residuals_2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
residuals_2	17,919	1.16e-11	.479876	-2.515588	2.142944

MR5 and MR6, the Gauss-Markov assumption and normality assumption, assume that the error term e is normally distributed for each value of x. However, referring to the central limit theorem we can assert that as sample size increases the distribution will assimilate a normal distribution, thus this assumption eventually will be true as N goes to infinity.

Q2) Data on U.S. gasoline consumption for the years 1953 to 2004 are given in Table F2.2 (http://pages.stern.nyu.edu/~wgreene/Text/Edition7/tablelist8new.htm). Note that the consumption data appear as total expenditure. To obtain the per capita quantity variable, divide GASEXP (total U.S. gas expenditure) by GASP (price index for gasoline) times Pop (U.S. population in thousands). The other variables do not need transformation.

a. Compute the multiple regression of per capita consumption of gasoline on per capita income, the price of gasoline, all the other prices and a time trend. Report all results. Do the signs of the estimates agree with your expectations?

```
. reg consum income gasp pnc puc ppt pd pn ps year
```

Source	SS	df	MS
Model	9.4872e+11	9	1.0541e+11
Residual	2.9283e+09	42	69721952.3
Total	9.5165e+11	51	1.8660e+10

Number of obs	=	52
F(9, 42)	=	1511.92
Prob > F	=	0.0000
R-squared	=	0.9969
Adj R-squared	=	0.9963
Root MSE	=	8350

consum	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
income	11.23384	3.967027	2.83	0.007	3.228059	19.23963
gasp	-802.8624	304.8832	-2.63	0.012	-1418.142	-187.5832
pnc	126.097	984.375	0.13	0.899	-1860.452	2112.646
puc	634.1018	373.2613	1.70	0.097	-119.1701	1387.374
ppt	219.8422	370.6409	0.59	0.556	-528.1415	967.8259
pd	-2570.917	910.6175	-2.82	0.007	-4408.618	-733.2167
pn	2409.452	965.5099	2.50	0.017	460.9743	4357.93
ps	-762.8231	612.8325	-1.24	0.220	-1999.569	473.923
year	4714.412	1086.97	4.34	0.000	2520.819	6908.006
_cons	-9212529	2084596	-4.42	0.000	-1.34e+07	-5005643

Worth mentioning are the statistically and practically significant opposite effects of the aggregate price index for consumer durables (pd) and consumer nondurables (pn). This indicates that when the price index of consumer durables - goods that are of long duration - goes up, the consumption of gasoline price decreases. This insight might suggest that if durables are necessity goods, and gasoline price is a luxury good. If the price of the necessity goods goes up, consumers may not be able to buy gasoline anymore. Thus whether or not this direction of coefficient makes sense or not depends on how we classify gas and durable goods.

If not mentioned, coefficients are statistically significant at an alpha level of 0.01.

consum	per capita consumption of gasoline	
income	per capita disposable income	(+) expectedly, the higher the income, the more individuals spend on gasoline
gasp	gasoline price index	(-) expectedly, the higher the price, the less individuals spend on gasoline *(significant at an alpha of 0.05)*
pnc	Price index for new cars	*not statistically significant*
puc	price index for used cars	*not statistically significant*
ppt	price index for public transportation	*not statistically significant*
pd	aggregate price index for consumer durables	(-) un,-expectedly, the higher the price of durables, the less individuals spend on gasoline
pn	aggregate price index for consumer nondurables	(-) un,-expectedly, the higher the price of durables, the less individuals spend on gasoline *(statistically significant at an alpha of 0.05)*
ps	aggregate price index for consumer services	not statistically significant
year	1953-2004	(+) expectedly, price increases with time

b. Test the hypothesis that at least in regard to demand for gasoline, consumers do not differentiate between changes in the prices of new and used cars.

```
. lincom pnc - puc

 ( 1)  pnc - puc = 0
```

consum	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-508.0048	999.1854	-0.51	0.614	-2524.443	1508.433

The lincom command in Stata takes the two coefficients of the chosen variables (prices of new cars and prices of used cars), creates a new coefficient that is their difference, and calculates whether or not this difference is close to zero via standard errors and t-statistic. From the reported t-statistic (-0.51) and p-value (0.614) we can conclude to reject the null hypothesis that consumers do not differentiate between changes in prices of new and used cars. The coefficients in the table above (126 for new cars and 634 for used cars), suggests that they more strongly react to a price change of used cars, but neither of the results are statistically significant.

c. Estimate the own price elasticity of demand, the income elasticity, and the cross-price elasticity with respect to changes in the price of public transportation. Do the computations at the 2004 point in the data.

Elasticities show how sensitive a variable is to the change of another variable. Here, we can show that individual consumption of gasoline is strongly sensitive to changes in income, and somewhat

If gas price increases by 1%, consumption decreases by roughly 10%. The other results are not statistically significant. (If they were: If public transportation increases by 1%, consumption would increase by roughly 5%, if income increases by 1%, consumption would increase by roughly 46%.)

```
. margins, eyex(gasp income ppt) at(year=2004)

Average marginal effects                        Number of obs     =         52
Model VCE    : OLS

Expression   : Linear prediction, predict()
ey/ex w.r.t. : income gasp ppt
at           : year            =        2004

------------------------------------------------------------------------------
             |            Delta-method
             |      ey/ex   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .4663318   .1948624     2.39   0.021     .0730835    .8595801
        gasp |  -.0984843   .0356338    -2.76   0.008    -.1703963   -.0265724
         ppt |   .0453609   .0770073     0.59   0.559    -.1100461    .2007679
------------------------------------------------------------------------------
```

d. Reestimate the regression in logarithms so that the coefficients are direct estimates of the elasticities. (Do not use the log of the time trend). How do your estimates compare with the results in the previous question? Which specification do you prefer?

These elasticity results look very different from the table above and I am unsure as to why. I would trust the (d) regression more, as I have more visibility into how the variables were transformed and computed.

```
. reg log_consum log_income log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps year
```

Source	SS	df	MS			
				Number of obs	=	52
				F(9, 42)	=	1351.37
Model	16.7315653	9	1.85906281	Prob > F	=	0.0000
Residual	.057778672	42	.001375683	R-squared	=	0.9966
				Adj R-squared	=	0.9958
Total	16.789344	51	.329202823	Root MSE	=	.03709

log_consum	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
log_income	1.045575	.308215	3.39	0.002	.423572	1.667578
log_gasp	.2152958	.0664869	3.24	0.002	.0811199	.3494718
log_pnc	-.0817842	.3286342	-0.25	0.805	-.7449948	.5814265
log_puc	-.5937312	.1048814	-5.66	0.000	-.8053904	-.3820721
log_ppt	.0927563	.1679697	0.55	0.584	-.2462202	.4317328
log_pd	2.286064	.3199229	7.15	0.000	1.640433	2.931694
log_pn	-1.267631	.3263016	-3.88	0.000	-1.926135	-.609128
log_ps	-1.338565	.434437	-3.08	0.004	-2.215294	-.4618354
year	.0818608	.0092494	8.85	0.000	.0631947	.1005269
_cons	-156.8193	16.14946	-9.71	0.000	-189.4103	-124.2284

e. Compute the simple correlations of the price variables. Would you conclude that multicollinearity is a "problem" for the regression in part a or part d?

Collinearity occurs when two or more variables are so strongly correlated with each other, that one is the function of another. For example, if we were to regress weight on height and 2xheight, the combination of these two explanatory variables would give rise to the problem of collinearity. This becomes a problem because the variation among them is not high enough to estimate precise coefficients. Mathematically, the origins is in the correlation matrix: if two columns are strongly correlated, their determinant is 0 and their matrix becomes non invertible.

```
. corr gasp pnc puc ppt pd pn ps
(obs=52)
```

	gasp	pnc	puc	ppt	pd	pn	ps
gasp	1.0000						
pnc	0.9361	1.0000					
puc	0.9228	0.9939	1.0000				
ppt	0.9270	0.9807	0.9824	1.0000			
pd	0.9389	0.9933	0.9878	0.9585	1.0000		
pn	0.9627	0.9885	0.9822	0.9899	0.9773	1.0000	
ps	0.9394	0.9785	0.9769	0.9975	0.9563	0.9936	1.0000

In this case, however, collinearity is not a problem. Even though prices are strongly correlated, with values above 0.9 for all of them, Stata would automatically return an error if they were too strongly correlated. To show this, I created a variable log_ps_2 which is equal to 2*log(ps). As seen in the table, Stata automatically omits one of them.

```
. gen log_ps_2 = 2*log(ps)

. reg log_consum log_income log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps year l
> og_ps_2
note: log_ps omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =   1351.37
       Model |  16.7315653         9  1.85906281   Prob > F        =    0.0000
    Residual |  .057778672        42  .001375683   R-squared       =    0.9966
-------------+----------------------------------   Adj R-squared   =    0.9958
       Total |   16.789344        51  .329202823   Root MSE        =    .03709

------------------------------------------------------------------------------
  log_consum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_income |   1.045575    .308215     3.39   0.002      .423572    1.667578
    log_gasp |   .2152958   .0664869     3.24   0.002     .0811199    .3494718
     log_pnc |  -.0817842   .3286342    -0.25   0.805    -.7449948    .5814265
     log_puc |  -.5937312   .1048814    -5.66   0.000    -.8053904   -.3820721
     log_ppt |   .0927563   .1679697     0.55   0.584    -.2462202    .4317328
      log_pd |   2.286064   .3199229     7.15   0.000     1.640433    2.931694
      log_pn |  -1.267631   .3263016    -3.88   0.000    -1.926135    -.609128
      log_ps |          0  (omitted)
        year |   .0818608   .0092494     8.85   0.000     .0631947    .1005269
    log_ps_2 |  -.6692824   .2172185    -3.08   0.004    -1.107647   -.2309177
       _cons |  -156.8193   16.14946    -9.71   0.000    -189.4103   -124.2284
------------------------------------------------------------------------------
```

f. Notice that the price index for gasoline is normalized to 100 in 2000, whereas the other price indices are anchored at 1983 (roughly). If you were to renormalize the indices so that they were all 100.00 in 2004, then how would the results of the regression in part a change? How would the results of the regression in part d change?

Original version (a)

```
. reg consum income gasp pnc puc ppt pd pn ps year

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =   1511.92
       Model |  9.4872e+11         9  1.0541e+11   Prob > F        =    0.0000
    Residual |  2.9283e+09        42  69721952.3   R-squared       =    0.9969
-------------+----------------------------------   Adj R-squared   =    0.9963
       Total |  9.5165e+11        51  1.8660e+10   Root MSE        =      8350

------------------------------------------------------------------------------
      consum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   11.23384   3.967027     2.83   0.007     3.228059    19.23963
        gasp |  -802.8624   304.8832    -2.63   0.012    -1418.142   -187.5832
         pnc |    126.097    984.375     0.13   0.899    -1860.452    2112.646
         puc |   634.1018   373.2613     1.70   0.097    -119.1701    1387.374
         ppt |   219.8422   370.6409     0.59   0.556    -528.1415    967.8259
          pd |  -2570.917   910.6175    -2.82   0.007    -4408.618   -733.2167
          pn |   2409.452   965.5099     2.50   0.017     460.9743     4357.93
          ps |  -762.8231   612.8325    -1.24   0.220    -1999.569     473.923
        year |   4714.412    1086.97     4.34   0.000     2520.819    6908.006
       _cons |   -9212529    2084596    -4.42   0.000    -1.34e+07    -5005643
------------------------------------------------------------------------------
```

Indexed version of (a)

There are some changes in the coefficients (e.g gas price from -802 to -994), but the p-value and t-statistic remained the exact same for each of the variables. In the end a regression tries to understand the comparative relationship between variables, and is able to disregard absolute values most of the time. Since a re-normalization/ re-indexing is giving the variables just a different starting point, thus shifts them up or down in their absolute values, it does not have a big effect on the relationships between them.

```
. reg consum income gasp_ind pnc_ind puc_ind ppt_ind pd_ind pn_ind ps_ind year

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =   1511.92
       Model |  9.4872e+11         9  1.0541e+11   Prob > F        =    0.0000
    Residual |  2.9283e+09        42  69721950.3   R-squared       =    0.9969
-------------+----------------------------------   Adj R-squared   =    0.9963
       Total |  9.5165e+11        51  1.8660e+10   Root MSE        =      8350

------------------------------------------------------------------------------
      consum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   11.23384   3.967024     2.83   0.007     3.228061    19.23962
    gasp_ind |  -994.7553   377.7533    -2.63   0.012    -1757.092   -232.4183
     pnc_ind |    168.843   1318.077     0.13   0.899    -2491.145    2828.831
     puc_ind |   845.2578   497.5573     1.70   0.097    -158.8534    1849.369
     ppt_ind |   459.6881   775.0103     0.59   0.556    -1104.346    2023.722
      pd_ind |  -2951.413   1045.388    -2.82   0.007    -5061.092   -841.7335
      pn_ind |   4149.078   1662.608     2.50   0.017        793.8    7504.357
      ps_ind |  -1699.568    1365.39    -1.24   0.220    -4455.037    1055.901
        year |   4714.412   1086.969     4.34   0.000      2520.82    6908.005
       _cons |   -9212529    2084595    -4.42   0.000    -1.34e+07    -5005645
------------------------------------------------------------------------------
```

Original version (d)

```
. reg log_consum log_income log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps year

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =   1351.37
       Model |  16.7315653         9  1.85906281   Prob > F        =    0.0000
    Residual |  .057778672        42  .001375683   R-squared       =    0.9966
-------------+----------------------------------   Adj R-squared   =    0.9958
       Total |   16.789344        51  .329202823   Root MSE        =    .03709

------------------------------------------------------------------------------
  log_consum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_income |   1.045575    .308215     3.39   0.002      .423572    1.667578
    log_gasp |   .2152958   .0664869     3.24   0.002     .0811199    .3494718
     log_pnc |  -.0817842   .3286342    -0.25   0.805    -.7449948    .5814265
     log_puc |  -.5937312   .1048814    -5.66   0.000    -.8053904   -.3820721
     log_ppt |   .0927563   .1679697     0.55   0.584    -.2462202    .4317328
      log_pd |   2.286064   .3199229     7.15   0.000     1.640433    2.931694
      log_pn |  -1.267631   .3263016    -3.88   0.000    -1.926135    -.609128
      log_ps |  -1.338565    .434437    -3.08   0.004    -2.215294   -.4618354
        year |   .0818608   .0092494     8.85   0.000     .0631947    .1005269
       _cons |  -156.8193   16.14946    -9.71   0.000    -189.4103   -124.2284
------------------------------------------------------------------------------
```

Indexed version of (d)

A logarithm is intended to make large values more easy to deal with, by making them smaller but keeping values in the same proportion. Because values are smaller, but their relationship is still the same, the difference between the original version of (d) and the indexed one becomes even smaller than before working in the log space. For example gas price changes from 0.2152958 to 0.2152948 - a miniscule difference.

```
. reg log_consum log_income log_gasp_ind log_pnc_ind log_puc_ind log_ppt_ind log_pd_ind
> log_pn_ind log_ps_ind year

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(9, 42)        =   1351.37
       Model |  16.7315653         9  1.85906281   Prob > F        =    0.0000
    Residual |  .057778676        42  .001375683   R-squared       =    0.9966
-------------+----------------------------------   Adj R-squared   =    0.9958
       Total |   16.789344        51  .329202823   Root MSE        =    .03709

------------------------------------------------------------------------------
  log_consum |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_income |   1.045578   .3082145     3.39   0.002     .4235755     1.66758
log_gasp_ind |   .2152948   .0664869     3.24   0.002     .0811189    .3494708
 log_pnc_ind |  -.0817832   .3286339    -0.25   0.805    -.7449932    .5814268
 log_puc_ind |  -.5937312   .1048813    -5.66   0.000    -.8053903   -.3820721
 log_ppt_ind |   .0927515   .1679696     0.55   0.584    -.2462249    .4317279
  log_pd_ind |   2.286064   .3199221     7.15   0.000     1.640435    2.931693
  log_pn_ind |  -1.267637   .3263014    -3.88   0.000     -1.92614   -.6091343
  log_ps_ind |  -1.338555   .4344363    -3.08   0.004    -2.215283   -.4618273
        year |   .0818607   .0092494     8.85   0.000     .0631946    .1005268
       _cons |   -158.345   16.27474    -9.73   0.000    -191.1887   -125.5012
------------------------------------------------------------------------------
```

Word Count: 1,200

Appendix

Regressions run with robust option

```
. reg $y_list $x1_list, robust

Linear regression                               Number of obs     =        109
                                                F(3, 105)         =      19.17
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3408
                                                Root MSE          =     .37906
```

logwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1283769	.0261456	4.91	0.000	.076535	.1802187
potexper	.0442788	.0112038	3.95	0.000	.0220637	.0664938
ability	-.1481151	.0446863	-3.31	0.001	-.2367198	-.0595103
_cons	.3292917	.3352545	0.98	0.328	-.335456	.9940393

```
. reg $y_list $x1_list $x2_list, robust

Linear regression                               Number of obs     =        109
                                                F(6, 102)         =      12.57
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3660
                                                Root MSE          =     .37716
```

logwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1359242	.0274855	4.95	0.000	.0814069	.1904416
potexper	.0475567	.0104694	4.54	0.000	.0267907	.0683226
ability	-.074606	.0927297	-0.80	0.423	-.258535	.109323
mothered	.0111151	.0444653	0.25	0.803	-.0770817	.0993118
fathered	-.0342395	.0185843	-1.84	0.068	-.0711014	.0026224
siblings	.0396161	.0618996	0.64	0.524	-.0831614	.1623936
_cons	.3815602	.673612	0.57	0.572	-.9545459	1.717666

Stata Code

```
* Assignment

* (1)
* data file fron Koop and Tobias (2004)
* http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataSets.htm
import delimited "/Users/annapauxberger/Documents/Minerva Academics III/SS154
Econometrics/Assignment 2/Koop-Tobias.csv", clear

* extract observations for first 15 individuals in the sample
drop if personid > 15

* define x and y lists
glob y_list  logwage
glob x1_list educ potexper ability
glob x2_list mothered fathered siblings

* line of best fit visualization
reg logwage educ
predict logwagehat
twoway(scatter logwage educ) (line logwagehat educ)

* (a) X1 regression
reg $y_list $x1_list
reg $y_list $x1_list, robust

* (b) X1 and X2 regression
reg $y_list $x1_list $x2_list
reg $y_list $x1_list $x2_list, robust


* (c) calculate R^2
disp e(mss) / (e(mss)+e(rss))

* without a constant
reg $y_list $x1_list $x2_list, noconstant
disp e(mss) / (e(mss)+e(rss))


* (d) adj. R^2
gen n = e(N)
gen r_2 = e(r2)
gen k = 6
disp 1-((1-r_2)*(n-1) / (n-k-1))


* (e) assumptions
quietly reg $y_list $x1_list
```

```
predict residuals, resid
scatter residuals logwage
summarize residuals

quietly reg $y_list $x1_list $x2_list
predict residuals_2, resid
scatter residuals_2 logwage
summarize residuals_2


* (2)
import delimited "/Users/annapauxberger/Documents/Minerva Academics III/SS154
Econometrics/Assignment 2/TableF2-2.csv", clear

* (a) regression
gen consum = gasexp/gasp*pop
reg consum income gasp pnc puc ppt pd pn ps year


* (b) test difference of new and used cars
lincom pnc - puc

* (c) estimate price elasticity
margins, eyex(gasp income ppt) at(year=2004)

* (d) regression in logarithms
gen log_consum = log(consum)
gen log_income = log(income)
gen log_gasp = log(gasp)
gen log_pnc = log(pnc)
gen log_puc = log(puc)
gen log_ppt = log(ppt)
gen log_pd = log(pd)
gen log_pn = log(pn)
gen log_ps = log(ps)

reg log_consum log_income log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps year
margins, eyex(gasp income ppt) at(year=2004)

* (e) correlations of price variables
corr gasp pnc puc ppt pd pn ps

gen log_ps_2 = 2*log(ps)
reg log_consum log_income log_gasp log_pnc log_puc log_ppt log_pd log_pn log_ps year log_ps_2

* (f) re-index to year 2004

gen gasp_ind = 100 * gasp / gasp[52]
```

```
gen pnc_ind = 100 * pnc / pnc[52]
gen puc_ind = 100 * puc / puc[52]
gen ppt_ind = 100 * ppt / ppt[52]
gen pd_ind = 100 * pd / pd[52]
gen pn_ind = 100 * pn / pn[52]
gen ps_ind = 100 * ps / ps[52]

reg consum income gasp_ind pnc_ind puc_ind ppt_ind pd_ind pn_ind ps_ind year

gen log_gasp_ind = log(gasp_ind)
gen log_pnc_ind = log(pnc_ind)
gen log_puc_ind = log(puc_ind)
gen log_ppt_ind = log(ppt_ind)
gen log_pd_ind = log(pd_ind)
gen log_pn_ind = log(pn_ind)
gen log_ps_ind = log(ps_ind)

reg log_consum log_income log_gasp_ind log_pnc_ind log_puc_ind log_ppt_ind log_pd_ind log_pn_ind
log_ps_ind year
```

Bibliography

Adjusted R2. (2018). Retrieved from https://www.statisticshowto.datasciencecentral.com/adjusted-r2/

Hill, R. C., Griffiths W. E., & Lim G. C. (Fourth edition). *Principles of econometrics* (pp. 130-156). Hoboken, New Jersey: John Wiley & Sons, Inc. Retrieved from https://www.amazon.com/Principles-Econometrics-R-Carter-Hill/dp/0470626739/ref=sr_1_1?s=books&ie=UTF8&qid=1530558152&sr=1-1&keywords=principles+of+econometrics